

# ARABIC USER SEARCH QUERY CORRECTION AND EXPANSION

T. Rachidi, M. Bouzoubaa, L. Elmortaji, B. Boussouab, and A. Bensaid  
 AlAkhawayn University in Ifrane,  
 PO. Box 1881, Ifrane 53000  
 Morocco  
 [T.Rachidi,A.Bensaid]@alakhawayn.ma

**ABSTRACT-** This paper describes correction and expansion techniques of multilingual search queries submitted to the Arabic-centred search engine Barq [1]. Key features of the correction technique are the use of **1.** Arabic language morphology, **2.** Arab speaker most common pronunciation and spelling mistakes in Arabic **3.** Arab speaker most common spelling mistakes in transliterated English, and **4.** Arabic language learners (as L2) spelling and pronunciation mistakes. The query expansion mechanisms also uses Arabic words roots and thesauri built automatically by a categorization engine, in order to achieve better recall. The preliminary results obtained for query correction show 92% of misspelled queries terms submitted are corrected. The preliminary results obtained for expansion show a promising 75% increase in recall, though qualitative evaluation for recall improvement is still underway.

**KEYWORDS** - Arabic query correction, Arabic query expansion, Arabic transliteration, Search engine

## I. INTRODUCTION

Today the web counts many Arabic portals and no less than seven (7) Arabic search engines [2-8]. Many of the portals comprise not only documents written exclusively in Arabic, but also multilingual documents that comprise both Arabic and other languages such as English or French. Scientific and technical reports, as well as Islam-related material are such documents. Many of the words in these documents are transliterated from English into Arabic, and from Arabic into English. The transliteration in either direction is different depending on the region of origin of the transliterator/writer. Table 1 and 2 give sample transliterations, from English to Arabic and from Arabic to English, encountered on the Arabic Web today.

| English word | Arabic transliteration North Africa | Arabic transliteration Middle East |
|--------------|-------------------------------------|------------------------------------|
| Television   | تلفزة                               | تلفاز                              |
| American     | الامريكي                            | الاميركي                           |

**Table 1.** Sample transliterations from English present in the Arabic Web today.

| Arabic word | English transliteration by North Africans | English transliteration in the Middle East |
|-------------|---|--|
| الجزيرة     | Eljazira                                  | Aljazera, Aljazeera, Aljazeera             |
| محمد        | Mohammed, Mohamed                         | Muhammad                                   |
| داود        | Dawoud                                    | Dawood                                     |

**Table 2.** Sample transliterations found in the Arabic Web.

Moreover, search engines that support Arabic are used by native Arab speakers from North Africa, the Middle East, as well as Muslims from all over the world, and Arabic language learners alike. Depending on the linguistic background of the user, query terms for the same search

are written differently affected by the L1 background. Table 3 gives sample queries and results obtained with leading search engines. These queries, although for the same thing, return different numbers of documents depending on term spelling, clearly indicating room for improvement in recall.

The inefficiencies advocate for a query correction mechanism that cater for all backgrounds. Such a mechanism must not only cope with the variants in transliterations from English to Arabic, and from Arabic to English that exist today on the Arabic Web, but also for Arabic spelling mistakes most commonly encountered with both Arab natives and learners of Arabic.

| Engine | Query string | # of docs. returned | Query string | # of docs. returned |
|--------|--------------|---------------------|--------------|---------------------|
| Hahooa | Mohammed     | 24                  | Muhammad     | 14                  |
| Google | Mohammed     | 1160000             | Muhammad     | 883000              |
| Ayna   | Mohammed     | 190                 | Muhammad     | 12                  |
| Engine | Query string | # of docs. returned | Query string | # of docs. returned |
| Hahooa | الامريكي     | 3                   | الاميركي     | 0                   |
| Google | الامريكي     | 29100               | الاميركي     | 20,500              |
| Ayna   | الامريكي     | 32                  | الاميركي     | 124                 |

**Table 3.** Results obtained with Arabic search engines.

Query correction applied to Web search engines is not new. Google for instance implements query correction through the "Did you mean hyperlinks". Nonetheless, Google does not implement Arabic specific features to query correction such as Arabic word morphology [9] or Arabic word root [10], nor does it implement user background specific query correction. As a matter of fact, as of October 2003, all engines in [2-8] offer basic keyword based search, without query correction.

In Arabic, as many as hundreds of words can derive from a single root [9]. Without retrieving documents containing all of the derivatives of the query terms, Arabic engines return a poor recall. Furthermore, a user's query is very often too short (less than two words on average [11]), ambiguous, and of the wrong granularity [12]. To overcome these problems query expansion techniques are used. Query expansion consists of expanding a query of few keywords to many related keywords, in order to give more recall to that query. Query expansion is crucial for Arabic search engines given the particular morphology of Arabic language.

Query expansion of Arabic terms requires language specific processing, namely word root extraction [10], and morphological analysis. Expanding Arabic words with related words necessitates automatic categorization of Arabic documents for concept building [14].

The work presented in this paper describes the techniques and the results obtained for both query expansion and query correction processes as implemented in the Barq [1]

search engine. Both processes take advantage of Arabic specific language features. Query correction, in particular, leverages of most common Arabic spelling mistakes by Arab natives and L1/2 Arabic learners. Query expansion uses word root extraction and thesauri built from automatic categorization of Arabic documents. Section 2 gives insight into techniques for word spelling correction, a key element in query correction, as well as insight into query expansion for the purpose of recall improvement. Section 3 presents the general query processing framework implemented in Barq. Section 4 and 5 present the techniques used for query correction and expansion in Barq respectively. Section 6 gives preliminary results obtained for both techniques.

## 2. QUERY CORRECTION AND EXPANSION

Query correction's main function is to correct user query terms. One such technique is statistical; it uses frequencies of digrams and trigrams to find possibly incorrect words. Words that contained digrams or trigrams of low frequency are judged possibly incorrect [15]. A more popular technique, of which we adopted a modified form in this work, uses dictionary lookups. Dictionary lookups produce a list of possible replacements for each term. Replacements are similar words present in the dictionary [16]. Three similarity metrics are commonly used: 1. the edit distance [17], 2. similarity of roots and morphological categories, and 3. pronunciation. Each of these metrics reflects the origins of errors. The edit distance between two strings, for instance, specifies how many basic editing operations are needed to convert one string into the other. The basic editing operations are insertion, deletion, change, and transposition (see Table 6).

In many languages, one can find words that have different spelling, but that are pronounced either identically, or in a similar way. Finding the words that are pronounced in a similar manner to a given word is usually done with transducers [19, 20]. For languages with highly regular pronunciation, such as Arabic, a simple solution that takes into account the possibility of replacing a pair of characters with a single character and *vice versa* can work efficiently. It is this solution that we adopt in our query correction for pronunciation errors. For example, the letters listed in Table 4 have similar pronunciation, and although there are native Arab speakers that distinguish between them, most Arab speakers do not, especially in North Africa.

| Similar Letters | Examples            |
|-----------------|---------------------|
| ط ت ث ة         | فتات vs. فتاة       |
| س ص             | أسطوانة vs. أسطوانة |
| ذ               | أستاذ vs. أستاذ     |
| ض ظ             | ضلم vs. ظلم         |

Table 4. letters with similar pronunciation and/or spelling.

Although there is a wealth of literature in the study of morphological, phonetical, and syntactical features of the Arabic language, these remain at the linguistic level. Very few researchers bridged the gap by leveraging of these features into computational systems. One has to say that Arabic enabled systems (operating systems in particular) have only become readily available in the past four (4)

years or so. To the best of our knowledge, the field of Arabic query correction in general, and word error correction in particular have not been studied at all.

Furthermore, in Arabic transliterated from other languages, some letters can or cannot be present depending on the origin of the transliterator, yielding different spelling but phonetically similar words. Sample of these are given in Table 5.

| Letters | English word | Arabic transliteration | Arabic Transliteration |
|---------|--------------|------------------------|------------------------|
| ا       | Paris        | باريس                  | بريس                   |
| ى       | Joshua       | جشوى                   | جشو                    |
| و       | Bush         | بوش                    | بش                     |
| ي       | kiwi         | كبوي                   | كبو                    |

Table 5. These special letters could be added or omitted yielding different spelling, but phoetically similar words.

| Mistake types          | Examples  |
|------------------------|---|
| <i>Morphology</i>      | استمحن vs. امتحن<br>صعلك vs. صعليك                        |
| <i>Edit</i>            | فتات vs. فتاة<br>الرسالة vs. البسالة<br>حرير vs. حرير     |
| <i>Grammar/ Syntax</i> | دخل المسلمون vs. دخل المسلمين<br>مقروأ vs. مقرو vs. مقروء |
| <i>Pronunciation</i>   | أسطوانة vs. أسطوانة<br>أستاذ vs. أستاذ<br>ضلم vs. ظلم     |

Table 6. Summary of Arabic error types.

As for query expansion, a variety of techniques are available for expanding user query [11]. Some reweigh query terms, and others do not. Query expansion with reweighing uses relevance feedback. It consists in increasing the weight of subsequent query terms appearing in relevant documents, and in decreasing the weight of query terms appearing in non-relevant documents. Various techniques such as vector feedback and probabilistic, for computing the weights are available in the literature. Query expansion without term reweighing on the other hand does not use relevance feedback. It uses thesauri to expand by adding terms to the original query. Added terms are looked up in thesauri that can be manually, automatically, or semi-automatically built. Many variants to query expansion without reweighing are in the literature. Some consist in not expanding high frequency query terms, others consist in using only top connected query terms etc. Others yet use expansion terms from relevant documents themselves. Although it has been shown that query expansion by related terms isn't as effective as query expansion by terms from relevant documents [23], Barq query expansion use relevance the latter. The main motivation behind this, is that this has not been proven for Arabic where as many as hundred (100) words can be derived from the one root.

## III. BARQ QUERY PROCESSING

Figure 1 shows the general view of query processing in Barq. As can be seen, query correction process does not interfere with query expansion and retrieval processes. The results of query correction in the form of hyperlinks to one or more queries are combined with the result page yielded



